原言語が異なる翻訳コーパスの定量的分析

白　　京 姫[†]　大竹　清敬[†]
Francis Bond[††]　山本　和英[†††]

日英パラレルコーパスにおける日本語と英語それぞれを原言語として翻訳した2つの韓国語コーパスを用いて，原言語が翻訳に及ぼす影響を調べた．コーパスにはATRのBTEC(162,308文)を使った．2つの韓国語コーパスは，日英パラレルコーパスからの翻訳であり，内容は一致している．それにも関らず，韓国語両コーパス間の同一文は3%以下であり，正書法が統一されていない点を考慮しても，同一または同一とみなせる文は全体の8.3%程度である．本研究では，両コーパスにおける違いを原言語の影響と予想し，分析した結果を報告する．

キーワード: 多言語コーパス，類似度，日本語，韓国語，英語，同一内容，同義表現，原言語の影響

Quantitative Analysis of Corpora with Different Source Languages

Kyonghee Paik[†] , Kiyonori Ohtake[†] , Francis Bond[††]
and Kazuhide Yamamoto[†††]

In order to investigate the effect of source language on translations, we examine two variants of a Korean translation corpus. The first variant consists of Korean translations of 162,308 Japanese sentences from the ATR BTEC (Basic Expression Text Corpus). The second variant was made by translating the English translations of the Japanese sentences into Korean. We show that the source language text has a large influence on the target text. Even after normalizing orthographic differences, fewer than 8.3% of the sentences in the two variants were identical. We describe in general which phenomena differ and then discuss how our analysis can be used in natural language processing.

KeyWords: *bilingual corpus, similarity score, Japanese, Korean, English, linguistic difference, same content*

1　背景

多言語コーパスが整備されていく過程で，ある言語への翻訳が複数の言語に基づいて行われる場合がある．たとえば，聖書の翻訳における日本語訳を考える際に，その原言語として様々な言語が存在する状況に類似している．原言語が英語とフランス語のような場合，それらからの日

† ATR音声言語コミュニケーション研究所, ATR Spoken Language Communication Research Laboratories
†† 日本電信電話株式会社 コミュニケーション科学基礎研究所, NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
††† 長岡技術科学大学 電気系, Department of Electrical Engineering, Nagaoka University of Technology

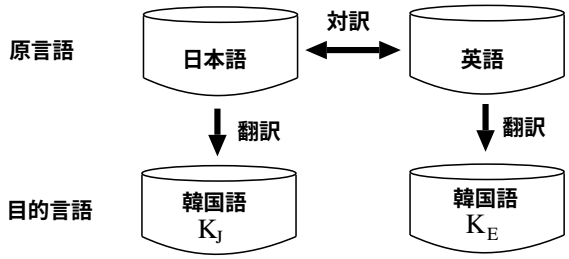

図 1　原言語が異なる韓国語コーパス

本語訳には，原言語の影響はほとんどないかもしれない．一方で，原言語として韓国語と英語のような対を考える場合，それらの原言語の違いは，翻訳に多大な影響を及ぼすと予想できる．

一般に，ある言語への翻訳が存在する場合，同一内容のものを別の言語から翻訳することは経済的理由から非常に少ない．たとえば，英語から日本語に翻訳された文書が存在する場合，同一内容の文書を韓国語から日本語に翻訳することは極めて稀である．まとまった量の文書を翻訳する場合，その可能性はさらに低くなる．そのため，原言語が異なる同一内容の大規模文書の翻訳は，人為的に作成されない限り入手は困難である．

一方で，原言語が翻訳に与える影響は確実に存在し，認識されている．ところが，これまで，原言語が翻訳に与える影響に関して，どのような現象がどの程度生じるのかについて詳細に調査した研究は存在しない．原言語によって生じる違いを詳細に研究することによって，人間と機械の双方にとってよりよい翻訳を得るための知見，知識が得られると考える．

そこで，本研究では，日本語と英語の対訳コーパスから日本語と英語を原言語として韓国語コーパスを作成し，翻訳における原言語の影響を考察する．各コーパスは，162,308 文から構成される．2つの韓国語コーパスと日英の対訳コーパスの関係を図1に示す．

これら2つの韓国語翻訳コーパスは原言語が日本語ならびに英語と大きく異なることから，それぞれ原言語の影響を受けたいくつかの特徴があり，両者は大きく異なる．本論文では，敬語表現，語彙選択，統語的差異，同義表現，表記のゆれ（正書法）の5つの言語現象の観点からそれらの違いを分析する．

周知のように英語は比較的固定された語順（SVO）を持ち，主語，目的語などが省略されない．反面，日本語は，述部が文末にくるが，それ以外の要素は，述部に対する関係を助詞などによって示すため，語順が柔軟である．さらに日本語では，文脈上明らかな主語，目的語などは明示されない．これらの点では，韓国語は英語より日本語に非常に近い言語である．このように，日本語と英語は，その構文構造が大きく異なる言語であり，語彙論的な観点からも，単語が与える意味や，その概念なども相当異なる．

(1) 韓国語: 그가 무례해서 화가 났다.
直訳: 彼が 無礼で 腹が 立った
英語: “His rudeness annoyed/bothered/upset me.”

たとえば，例 (1) に示した韓国語と英語の 2 つの文 (Lee 1999) は，同じ内容を表しているが，韓国語は複文構造を，英語は単文構造をとっている．これは，英語と日本語の間の翻訳についても言えることであるが，一方の言語において自然な表現を翻訳する場合，目的言語における自然な構文構造が，原言語のそれとは大きく異なる場合がある．しかしながら，翻訳が理想的な状況で行われるとは限らず，原言語の構文構造をそのままに，単語や，句を目的言語の該当表現へ変換することによって翻訳する場合もある．

したがって，原言語が日本語と英語のように大きく異なる言語からの韓国語翻訳文は，その原言語に大きく影響されると予想する．構文構造が大きく異なる言語間の翻訳において，目的言語における自然な構文へ翻訳することは，人間にとっても機械にとっても当然負担がかかる．以下に示す日本語と英語から韓国語への翻訳は，原言語の違いが翻訳に与える影響をよく示している．

(2) a. このケーブルカーに乗れば，ホテルに行くことができます。(原文)
訳: “이 케이블카를 타면 호텔에 갈 수 있습니다.”
この ケーブルカーに 乗れば ホテルに 行く ことが できます。
b. This cable car will take you to the hotel. (原文)
訳: “케이블카가 호텔에 데려다 줄 겁니다.”
ケーブルカーが ホテルに 連れて あげる でしょう。

例 (2a) の韓国語訳は，和文の構造をそのまま用いて翻訳されている反面，(2b) の韓国語訳は英文の構造に影響されている．

訳の自然さに関しては，日本語の構造の影響を受けている (2a) が (2b) に比べて非常に良い．この例から，より自然な文へ翻訳するために構文構造の大きな変更が必要な場合，そのような変更が行われず，原言語に大きく影響された翻訳が数多く存在していると予想する．原言語の違いが翻訳に差をもたらす事実は，認識されてはいても，これまで詳細に検討されたことはなかった．本研究では，両コーパスの分析を通して翻訳における原言語の影響を計量的に示し，このような異質なコーパスを機械翻訳および他の自然言語処理の分野にどのように応用できるかについて考察する．

2 原言語が異なるコーパスの比較

本論文では，ATR 旅行会話基本表現集（BTEC）を用いる．BTEC は旅行会話に必要とされる話題（たとえば，買物，ホテル・レストラン予約など）をはじめ，旅行に関する様々な話

題を網羅している (Takezawa, Shirai, and Ooyama 2001)．BTEC は当初，日本語と英語の対訳コーパスの収集から開始されたが，日本語または英語を原言語として他言語へ翻訳し，拡充してきた．本論文では，日本語から韓国語へ翻訳されたコーパスを K_J，英語から韓国語へ翻訳されたコーパスを K_E と表記する．また，BTEC は文単位の対応がとれた多言語コーパスである．使用する BTEC の構成を表1 に示す．また，使用する BTEC の一部の例文を付録 A に示す．

表 1　BTEC の構成

	日本語	中国語	英語	K_J	K_E
のべ	162,320	162,320	162,320	162,320	162,308
異なり	102,247	96,309	97,326	103,051	92,816
重複割合	37.0%	40.7%	40.0%	36.5%	42.8%

本論文では，K_J と K_E の2つのコーパスを比較し，それぞれの性質を詳しく調べる．まずは，2つのコーパスの間の類似度の分析から始め，次の節ではいくつかの言語現象についてより詳しく分析する．

2.1　類似度を用いたコーパス比較

まず，2つのコーパスが表層的にどの程度類似しているのかを編集距離に基づく類似度によって求めた．類似度を求めるプログラムは Perl 言語と `String::Similarity` モジュール (Lehmann 2000) を用いて作成した．このモジュールは，Myers による方法 (Myers 1986) によって，編集距離を求め，その値に基づき類似度を与える．具体的には2つの文字列 S_1, S_2 の類似度 sim は次の式によって与えられる．

$$sim(S_1, S_2) = 1 - \frac{Ins + Del}{|S_1| + |S_2|} \tag{1}$$

ここで，$|S|$ は文字列 S の長さに対応し，Ins と Del は2つの文字列を同一にするために必要な最小の編集操作 (挿入と削除) のそれぞれの回数を示している．たとえば，文字列 S_1 =“foo”, S_2 =“fou” の場合，S_1 を S_2 と同一とするための最小の編集操作は S_1 の最後の “o” を削除し，“u” を挿入することである．したがって，$Del = 1$, $Ins = 1$ となり，この場合は，

$$sim(S_1, S_2) = 1 - \frac{1+1}{3+3} = 0.6667$$

となる．一方，S_1 =“foo”, S_2 =“bar” と2つの文字列が全く異なる場合，S_1 を S_2 と同一にするためには S_1 の全ての文字列を削除し，S_2 と同一の文字列を挿入することになるので，$Del = 3$, $Ins = 3$ となり，この場合は，

$$sim(S_1, S_2) = 1 - \frac{3+3}{3+3} = 0$$

となる．したがって，simは全く異なる文字列に対しては0を，同一の文字列には類似度1を与える．具体例を以下に示す[1]．

(3) 100　　달러　예금하고　싶은데요.　(K_E)
　　백　　달러　예금하고　싶은데요.　(K_J)　　(類似度: $0.8750{=}1-\frac{1+3}{17+15}$)
　　100(百)　ドル　貯金し　たいです。

まず，K_J と K_E すべての翻訳対毎に類似度sim を求めた．そして，類似度0から1までを0.1刻みで10のクラスに分類し，それぞれを類似度クラス0から9まで(たとえば，類似度クラス1は類似度0.1以上0.2未満を示す)とした．この類似度を用いた集計結果を表2に示す．

表 2　類似度によるコーパスの比較結果

類似度	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6
類似度クラス	0	1	2	3	4	5
のべ	100	1,910	11,006	23,126	33,755	34,888
異なり	58	1,243	7,876	19,351	29,053	30,149

類似度	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0	Total
類似度クラス	6	7	8	9	
のべ	28,083	17,400	7,693	4,347	162,308
異なり	24,382	14,946	6,434	3,037	136,529

表2から分かることは，全ての異なり対のうち，全く異なる表現と判断された文が0.1%(58/136,529)以下で，ほぼ同一である文が3%以下(3037/136,529)である．使用したコーパスにおいて正書法が一貫してないところがあり，同一と判断されない文が多数存在する．この点を考慮すると，同一とみなせる文は約8.3%となる．

文献(Culy and Riehemann 2003)では，BLEU(Papineni, Roukos, Ward, and Zhu 2001)に代表される統計的な翻訳評価指標に関する実験を通して，一つのテキストに対して実に様々な訳が可能であることを議論している．この議論から，K_J と K_E が，原言語が異なるとはいえ同一の文が約8%程度であるということは，それほど不自然ではない．しかし，彼らが用いたテキストは，聖書と文学作品である．また，すでに翻訳が存在する場合に，翻訳者がその翻訳と差をつけようとする可能性があることや，原言語の影響に関してそこでは直接議論されていない．

3　諸言語現象における原言語の影響

原言語が翻訳に与える影響を調べるために，いくつかの言語現象に着目して，それぞれのコーパスを調査した．分析した言語現象は原言語の差が明確に現れる敬語表現，語彙選択（漢語，外来語），統語的差異（ゼロ代名詞，助数詞），文レベルの同義表現（意訳），表記のゆれ

1 韓国語では，文節ごとにスペースを置くが，スペースも編集距離の計算対象となる．

表 3　K_J と K_E における言語現象とその頻度

類似度クラス	0	1	2	3	4	5	6	7	8	9
無作為抽出数	58	12	78	193	290	301	243	149	64	30
敬語 (K_J)	4	1	10	16	82	101	76	32	11	2
敬語 (K_E)	2	0	8	15	20	32	31	9	2	0
ゼロ代名詞 (K_J)	0	0	3	8	23	28	22	13	4	1
ゼロ代名詞 (K_E)	0	0	2	5	7	15	8	9	1	1
漢語 (K_J)	0	0	7	23	39	54	22	19	4	0
漢語 (K_E)	0	0	2	10	25	35	19	12	0	0
外来語 (K_J)	0	0	5	16	14	15	7	6	0	0
外来語 (K_E)	0	0	0	7	11	9	11	3	0	0

(正書法) である．これらの言語現象のコーパス中の出現頻度を調査した．

実際の調査は，表2に示した類似度分布のうち，類似度クラス1から9までの範囲からそれぞれ1%にあたる文 (ペア) を無作為抽出し，それぞれの文がどのような現象を含んでいるかを調査した．類似度クラス0についてはすべての文を調べた．

まず，抽出された文のうち，K_J と K_E それぞれの文に敬語，ゼロ代名詞，漢語，外来語がどの程度含まれているかを調べた．集計結果を表3に示す．各項目は，無作為抽出した文のうち，それぞれの現象が含まれていた文数を示している．

次に対象とした対それぞれについて，文全体の対応，訳語選択における違い，統語的な違い，正書法の違いがそれぞれどの程度含まれているかを調査した．調査結果を表4に示す．抽出数などは表3と同一である．それぞれの項目は，各現象が含まれていた文数を示している．

3.1　敬語表現

話し手の発話の中で聞き手または話題にあがった第三者に対して，敬意を払っていることを示す表現が東洋の言語には多く見られる．韓国語ならびに日本語にもそのような敬語法が存在する．韓国語には，多段階の敬語レベルがあり，話者，聴者，第三者との間に社会的地位，年齢，グループ，親族関係，新密度などを考慮し使い分ける (Sohn 1999)．以下，韓国語の敬語体系を簡単に紹介する（日本語訳を“ ”で示す）．

話し手と聞き手との関係　김　교수(普通) vs 김　교수님 (敬語) “キム教授”

名詞　밥 (普通) vs 진지 (敬語) “ご飯/食事”

代名詞　나 (普通) vs 저 (謙譲) “私”

表 4　同義表現の種別とその頻度

		類似度クラス									
類型	現象	0	1	2	3	4	5	6	7	8	9
文全体	同一	0	0	0	0	0	0	0	3	0	18
	意訳	41	6	26	36	13	8	4	4	0	0
	誤訳	12	3	8	6	7	1	2	2	6	0
訳語選択	名詞	0	0	30	98	110	98	70	23	15	1
	動詞	0	0	32	112	193	186	88	37	11	2
	疑問詞	0	0	2	11	9	10	2	4	1	0
	その他	7	1	17	68	115	89	40	16	1	1
統語	助数詞	0	0	2	5	11	6	2	0	0	0
	その他	0	1	20	71	109	156	96	34	15	5
正書法	表記のゆれ	0	1	0	1	7	7	0	0	0	0
	数字	5	0	6	14	18	20	25	10	0	0
無作為抽出数		58	12	78	193	290	301	243	149	64	30

助詞　가, 이, 는 (普通) vs 께서, 께서는 (敬語) “が, は”

動詞　자다 (普通) vs 주무시다 (敬語) “寝る/眠る vs おやすみになる”

動詞接辞　ϕ (普通) vs -시 (主格敬語) （日本語には存在しない）

ϕ (普通) vs -ㅂ, -습 (相手敬語[2])“-られる”

文末表現　-(스)ㅂ니다 (尊敬), -어요/-아요 (丁寧), -소,오(対等[3]), -네(親密[4]), -어/-아(懇意[5]), -다 (疎遠[6]) “だ，です”

日本語の敬語形式は尊敬の度合い，話者，状況によって異なる．韓国語と同じように，接辞を付加する，異なる語彙を用いるなどによって敬語形式を構成する．以下，上記の韓国語の敬語法の例に対する日本語の敬語の種類を簡単に紹介する (Kaiser, Ichikawa, Kobayashi, and Yamamoto 2001).

名詞　:人 vs 方

名詞接頭辞　ϕ vs 御（お，ご，み）（例：お財布，お塩，ご都合，み心）

代名詞　私 (普通) vs 私（わたくし）(謙譲)

2 話しかける相手に対して用いる敬語.
3 相手にふさわしい尊敬を表し，相手と自分の間に心理的距離を維持したい時に用いられる.
4 対等，目下の者に対して使用する.
5 対等な間柄や目下の者に対して使われるが，-소,오 や -네に比べ柔らかな印象を与え，親近感を伝える場合に使われる.
6 この文末表現は前2つの “-네, -어/-아” と同程度に丁寧だが，親密さを欠いた表現となる.

形容詞　：いい/良い vs 宜しい, 暑い vs お暑い

動詞　食べる vs 召しあがる

動詞句　：主語：尊敬：(お/ご)+動詞の連用形+になる

話し手：謙譲：(お/ご)+動詞の連用形+する

文末表現　：た（普通），です，ます（丁寧）

上記のように韓国語と日本語は複雑な敬語体系を持っている．その反面，英語やその他のインド・ヨーロッパ諸語では代名詞形や敬称（たとえば，ドイツ語2人称 du に対して敬称である Sie など）に僅かに敬語の体系が残っているにすぎない．したがって，表現を体系的に変化させて尊敬を表すことができず，ほとんどの場合，全く異なる表現を用いることになる．表5に韓国語，日本語，英語の差を示す．

表 5　韓国語，日本語，英語における敬語の差

	韓国語	日本語	英語
普通	조금 기다린다.	ちょっと待つ.	"ϕ wait a moment. "
尊敬語	조금 기다리시겠습니까?	少し **お**待ち**になりますか**.	"Would you **mind** wait**ing** for a moment?"
謙譲語	조금 기다리겠습니다.	少し**お**待ち**します**.	"**If it is okay,** I'd **like to** wait for a moment."
丁寧語	드시다, 잡수시다	召しあがる	dine (vs. eat)

表5から三つの言語がそれぞれ独特な敬語法を用いていることが分かる．韓国語の謙譲語は日本語のような文法範疇を持たないことと，談話レベルにおいて文末の語尾が非常に多様であるという点が異なる．しかし，一方で，韓国語，日本語それぞれにおいて敬語の用法に規則性があることも示されている．英語の場合は，状況に応じて語句を挿入するなどして文を換え，敬意を表したり，自分をへりくだって表現する．表5では，英語において太字で示された表現に直接的に対応する表現は韓国語と日本語には存在しない．

以下，K_E と K_J における敬語表現の代表的な例を示す．

(4)　a. 호텔　내에　약국이　**있나요?:**　K_E

ホテル　内に　薬局が　ありますか。

"Do you have a drugstore in this hotel?" (原文)

b. 이　호텔에　약국은　**있습니까?:**　K_J

この　ホテルに　薬局は　ありますか。

"このホテルにドラックストアは ありますか。" (原文)　(類似度: 0.6207)

この例では，K_J の文がより丁寧に翻訳されている．一方，K_E の原文 "Do you have a drugstore in this hotel?" は丁寧さが明らかではなく，どちらにも訳せる．もし，例(4a)の原文が "Excuse me, would you please tell me if there is a drugstore in this hotel?" であれば，より丁

寧な表現に訳されたかも知れない．韓国語における語尾の 「-요」 は 「-ㅂ니까, -습니까」 に比べると丁寧ではない．これに直接対応する表現は日本語には存在しない．実際の会話においては (4a) は (4b) に比べ丁寧ではないが，日常会話ではよく使われる表現である．例文 (4b) は丁寧であるが，かしこまった場面により相応しいと言える．表 3から，K_Jでは，類似度に関係なく敬語表現を多用していることがわかり，原言語の違いが敬語表現の差異を形成している．

表6は，敬語（丁寧さ）に関係する文末表現が，K_JとK_Eでどの程度使用されているかを示したものである．敬語に関連する文末表現は表6に示した以外にもある（たとえば，-소, -오, -네, -어/-아 （親密な関係を表すもの）と -다 (普通)）が，本研究で用いたコーパスにはほとんど出現しない．本研究で用いたコーパスは旅行会話に関するものであり，そのほとんどがサービスを提供する側とされる側でのかしこまった会話で占められる．したがって，本研究で用いたコーパスには，親密な関係を示すようなくだけた表現はほとんど含まれない．表6の結果は，敬語に関する文末表現における原言語の差異を顕著に表している．また，参考までに，韓国語の文末表現と直接的に対応するわけではないが，日本語コーパスにおける文末表現を集計し，同じく表6にまとめた．

韓国語の文末表現が日本語のそれと直接的に対応しないというのは，日本語の “です/ます” は韓国語における “-ㅂ니다, -습니다” と “요” の両方に対応する可能性があるからである．つまり，韓国語の方が敬語のレベルがより細かいため，日本語の形式と直接的に対応しない．また，韓国語の “-ㅂ니다, -습니다” は敬語のレベルが最高丁寧と説明されるが，日本語にはこれらの語に対応する敬語形式は存在せず，“でしょうか” や “ましょうか” といった形式でより丁寧な表現にする．このような場合はほとんど「-ㅂ니까, -습니까」で翻訳されている．

表 6 韓国語と日本語コーパスにおける敬語文末表現

	コーパス			
文末表現	K_J	K_E	日本語	
ㅂ니다.（最高丁寧）	**33,351**	23,316	です．/ます．	18773/20373
ㅂ니까?（最高丁寧）	**34,872**	9,970	ですか．/ますか．	25759/27788
요.（普通丁寧）	21,922	**33,617**	だ．	806
요?（普通丁寧）	3,082	**25,481**	か．	225

3.2　語彙選択

韓国語と日本語は文法の面でも類似しているが，語彙的な面においても非常に近いと言える．ここでは，漢語と外来語に関して分析する．

漢語

漢字は中国をはじめ，日本，韓国等で使われている．日本語と韓国語における漢語は，共通のものが多く，韓国語における漢語の約七割が日本語にも存在すると言われている (渡辺・鈴木 1981)．Sohn によると，現代の韓国語の語彙は，純粋な韓国語 (35%)，中国語起源の韓国語 (60%)，そして外来語 (5%) で構成されている (Sohn 1999)．中国語起源の語彙はさらに，3 通りに分けられる．以下は文献 (張 2000) を参考にした．

中国語から輸入したもの　自然，天地，愛国，学院，英文，議事堂など
韓国語で作られたもの　便紙（“手紙”），福徳房 （“不動産屋”）など
日本語から輸入したもの　飛行機，旅行，英語，開戦，改良，会員など

さらに複雑になると，日本で作られた漢語が中国に逆輸入され（例，消防車，消化器，飛行機，旅行など），その漢語が韓国に輸入されたものもある．流入の経路が複雑で起源がはっきりしないものもあるが，日本語と韓国語は漢語の 7 割りが共通であることは非常に興味深い現象である．

実際に，表 3 でも，K_J のほうで漢語がよく使われている．以下に例をあげる．

(5) a. 리넨　제품　코너는　어디예요?:　K_E
リネン　製品　コーナーは　どこですか。
“Where is the **linens section**?”（原文）

b. 침구　매장은　어디입니까?:　K_J
寝具　売り場は　どこですか。
“**寝具売り場**はどこですか。”（原文）　(類似度: 0.3571)

(6) a. 보스턴　**가는**　버스는　어디에서 탑니까?:　K_E
ボストン　行きの　バスは　どこで　乗りますか。
“Where can I catch a bus **to go** to Boston?”（原文）

b. 보스톤　**행**　버스는　어디에서 탈　수　있습니까?:　K_J
ボストン　行き　バスは　どこで　乗る　ことが　できますか。
“ボストン**行き**のバスにはどこで乗れますか。”（原文）　(類似度: 0.7272)

表 3 で示したように漢語は K_J でより頻繁に使用されている．さらに，表 4 に示した訳語選択の項目において類似度クラス 3 から 7 まで名詞の訳語選択に違いが多く存在していることから，名詞訳語選択の違いの多くは漢語の使用にあることが推察できる．

外来語

外来語は外国から来た語（ただし，漢語を除く）であるが，その原言語は英語だけではない．韓国語と日本語には，英語，ポルトガル語，フランス語，ドイツ語など様々な国からの外来語が多く存在する．その割合を示した研究がある．日本語で書かれた90種類の雑誌から外来語を調査した結果，合計2,964個の外来語が収集された．その中で英語からの外来語は2,395語 (約80%) という圧倒的な数を示す (Shibatani 1990)．本研究の調査でも，ほとんどが英語からの外来語である．

表 3に示した結果から，外来語の量に関しては，K_J, K_E 両コーパスにおいて大きな差はなかった (K_J:4.6%(63/1358), K_E:3.0%(41/1358)) が，その性質は少し異なる．K_E で使われる外来語の中では “track, main dining room, avenue, check, coupon, dark brown, rent, seat, golf round” といった韓国語の外来語としてはなじみが薄い語も含まれている．反面，K_J で使われる外来語は “size, center, tour, ticket, economy car, room service, beer can, cream, family, restaurant, platform, curl, course, music, service, counter, play” のように和製英語および韓国語における外来語としてよく用いられる語が頻繁に表れる．韓国語の外来語には漢語と同様に日本から輸入されたものが数多くある．それゆえ，K_J で使われた外来語がそのまま異和感なく用いられている．このような事情もあって，日本語から韓国語に翻訳された場合，韓国語を母語とする者にとって親しみのある外来語が用いられる．しかし，英語からの翻訳においては，英語の単語をそのまま翻字した (5a) の “리넨(リネン)” のような外来語が含まれる．この現象は，原言語が訳語選択に大きく影響していることを示している． また，(5a), (5b) の例に示した場合のように “リネン製品” と “寝具” のような訳は広い意味では類似カテゴリーとして考えられるが，それらが与える意味は若干異なる．ひとつの原言語からの翻訳のみでは，このような2つの単語の類似関係はなかなか得られない．異なる原言語からの翻訳において，このような意味のずれが派生してくることは注目すべき点である．

日本語と英語を原言語とする韓国語の翻訳においては，漢語と外来語の用いられ方に原言語の影響が明確に表れる．

3.3 統語的差異

言語類型論的な観点からみると，韓国語は日本語に非常に近いが英語とはかなりへだたりがある．表4の統語の「その他」には格助詞の有無，文体の変化（能動対受動），構文の違いなどがある場合を計数した．表4から全ての類似度クラスにおいて統語的な違いがあることがよく分かる．本節では，特にゼロ代名詞と助数詞の2つの統語的な差異に焦点を当てる．

ゼロ代名詞

表3からK_Jでゼロ代名詞が頻繁に用いられていることが分かる．これは，1節で述べたように日本語では，文脈上明白な成分は明示されない事実を反映した結果である．以下に，文脈上明白な成分が明示されない場合の例を示す．

(7) a. 제　친구　집에　머물　예정입니다.: K_E
私の　友達　家に　泊る　予定です.
“I am planning to stay at my friend’s house.”（原文）

b. ϕ_{adn} 친구　집에　묵습니다.:　K_J
友達　家に　泊ります.
“ϕ 友達の家に滞在予定です．”（原文）　(類似度: 0.6429)

(8) a. ϕ_{subj} 분명히　이　비행기를　리컨펌했는데요.:　K_E
確かに　この　飛行機を　リコンファームしました.
“I’m sure I reconfirmed this flight. ”（原文）

b. ϕ_{subj} 확실히　ϕ_{obj} 재확인　했습니다.:　K_J
きちんと　再確認　しました.
“きちんとϕ_{subj} ϕ_{obj} リコンファームをしました．”（原文）　(類似度: 0.3125)

韓国語でも日本語と同様に文脈上明らかな成分は明示されない．上記の例は，原言語が英語の場合は，文脈上明らかな成分も明示される傾向があることを示している．このことは表3で示したゼロ代名詞の項目からも確認できる．

助数詞

助数詞は東アジアの言語で多く使われる．韓国語と日本語は英語のように直接物を数えることはできず，助数詞を数と共に用いる．英語の非加算名詞の数量表現に助数詞が使われるが（たとえば，two pieces of information），加算名詞の場合は助数詞を用いない(たとえば，one apple, two dogs)．両言語とも約300個の助数詞を持っているといわれるが，実際に日常生活で使われている典型的な助数詞の数は韓国語の場合，50個であり (Unterbeck 1994)，日本語の場合，おおよそ30個から80個である (Downing 1996) といわれる．以下，日本語，韓国語，英語における一例を示す．

(9) a. 日本語: 2**匹**の犬，何**頭**の牛[7]

b. 韓国語: 2**마리**의 개, 몇 **마리**의 소

7 日本語では対象とする動物により異なる助数詞が使われるが，韓国語ではほとんど「마리」を用いる．

c. 英語: “two dogs”, “how many cows”

2つのコーパスにおいて用いられる助数詞の種類ならびにその頻度を計数した結果を表7に示す．コーパス上，助数詞の面でも韓国語と日本語は類似している．表7から，K_J の方が助数詞を多く使用していることがわかる一方，その種類においては逆転しているという興味深い現象がおきている．つまり，用いられる助数詞の種類は K_E の方が12個も多い．この理由は，原言語において助数詞を明示するか否かによるところが大きいと考える．

表 7 K_J と K_E における助数詞の使用頻度

助数詞	K_E	K_J
種類	310	298
頻度	7,076	8,307

K_E の原言語コーパス（英語）においては，助数詞はまれにしか使われておらず，その多くは非加算名詞を数える時に限られている．したがって助数詞が明示されていない原言語文（英文)から，助数詞を用いる言語の文（韓国語文）に翻訳する時（たとえば，例文（10）を参照)，どのような助数詞を用いるかは，翻訳者に一任される．逆に助数詞が明示された原言語文（日本語文）から同じく助数詞を明示する目的言語文（韓国語文）へ翻訳する場合は，明示された助数詞の影響を受け，限定される．

たとえば，英語から韓国語への翻訳では，例文（10）のように，英語の “some water” を韓国語に訳す時 “잔 (杯)” または “컵 (コップ)” どちらを用いても訳せる．一方，日本語から韓国語への翻訳は “杯” は “잔”，“コップ” は “컵” のように訳が限定される傾向がある．例文（11）は日本語から韓国語への翻訳の例であり，原言語の助数詞を忠実に訳した場合を示している．

(10) 물 한 **잔/컵** 주세요.: K_E
お水 一 杯/コップ ください.
“Please give me some water.” （原文）

(11) 몇 **번**째 역에서 갈아탑니까?: K_J
何 番目 駅で 乗り換えますか.
“何**番**目の駅で乗り換えますか．” (原文）

3.4 同義表現の分類

本研究では，原言語が異なる2つの韓国語コーパスを比較し，様々な面について分析した．分析対象のうち約92%の文に不一致表現が含まれる．しかも，それらの文は同義表現であることから，そこにどのような差異が存在するかを調べることによって異表記同義表現を抽出できる．

表4に示した結果から，全体的に，名詞または動詞の訳語の違いが非常に多いことがわかる．

一方で，誤訳の多さも目立つ．これは，BTECの翻訳が，文脈をほとんど与えられず，一文単位で翻訳されていることに一因がある．類似度がかなり低い(類似度クラス0から2)文を分析すると，文単位の異表記同義表現が数多く得られると予想する．ただし，類似度の性質から，文が短い場合は，文単位の異表記同義表現であっても類似度が大きくなる傾向がある(例14aを参照)．以下に例をあげる．

(12) a. 욕실 딸린 싱글룸을 예약하고 싶습니다.: K_E
　　バス 付きの シングルルームを 予約し たいです
　　“A single with bath, please.”（原文）

　　b. 욕조 딸린 싱글 룸을 부탁합니다.: K_J
　　バス 付き シングル ルームを 付託する(お願いします).
　　“バス付きのシングル部屋を願います．”（原文）　(類似度：0.6315)

(13) a. 저는 야행성이에요.: K_E
　　私は 夜行性です.
　　“I’m a night owl. ”（原文）

　　b. 나는 밤 늦게까지 잠을 안 잡니다.: K_J
　　私は 夜 遅くまで 眠りを ない 眠ります.
　　“私は夜ふかしです．”（原文）　(類似度: 0.2069)

(14) a. 밥 생각이 없어요: K_E
　　ご飯 思いが ないです.
　　“I do not have any appetite.”（原文）

　　b. 식욕이 없습니다.: K_J
　　食欲が ありません.
　　“食欲がありません．”（原文）　(類似度: 0.4210)

例文(12a)と(12b)は類似度が高く，その差異を規則的に表現することが可能である．一方，例文(13a)，(13b)と(14a)，(14b)の間の差異は規則的なものではなく，熟語・慣用表現に属するものであり，規則的に変換することは非現実的である．これらは原言語である英語からの直訳ができず，熟語的，慣用的な表現を用いて翻訳されているからである．比較的単純な換言規則を用意し，それらを適用するだけでは，このような同義表現へ相互に換言することは困難である(Ohtake and Yamamoto 2001)．一方で，このような同義表現は類似度が低くなるにつれて増える傾向にあり，類似度を適切に用いることによって，比較的容易にこれらの同義表現を収集することができる．

3.5 正書法

K_J と K_E は別々に翻訳されたものであり，外来語や固有名詞の書き方と数字の書き方等に一貫性がない．たとえば，英語の “hostess” を含む文の翻訳をみると，K_J では “호스테스 (ホステス)”，K_E では “호스티스 (ホスティス)” のように表記されている．地名も同様である（たとえば，피카딜리 (pikadilli) と 피카디리 (pikadili)，애너하임 (aeneohaim) と에너하임 (eneohaim) など）．これらは，日本語における片仮名表記のゆれに該当する．また，数字表現では，“6시（時）40분（分）” と “여섯시 사십분” のように異なる表記が用いられる場合がある．つまり，アラビア数字 (1,2,3,...) で記述するか韓国語 (일，이，삼，...) で記述するかの違いである．原理的には，数字を漢字で記述することもできるが，この現象は，本研究で用いた2つのコーパスには存在しなかった．したがって，このコーパスを用いて，同義表現獲得，あるいは換言規則の抽出などを行う際には，表記の統一を考慮する必要がある．なお，このような異表記はコーパス全体の7%の文に存在する．

また，分かち書きの問題もある．日本語は分かち書きをしないので，この問題は存在しない．しかし，韓国語では「싱글룸」（シングルルーム）と「싱글␣룸」（シングル␣ルーム，例12aと12bを参照）のように分かち書きがゆれる場合がある．また，我々が評価した文には含まれていなかったが，コーパス全般において句読点の使用方法，特にクエスチョンマークの使用方法に関しても若干のゆれがある．

これらの正書法に関する違いを統一すると，同一と見なされる文は倍以上に増え，コーパス全体の8.3%になる．正書法に関する違いは，重要ではないもののコーパス中のいたるところに存在する．したがって，コーパスを計算機で処理する際に，正書法に関する違いをどのように扱うかは，非常に重要である．

4 2つのコーパスと自然言語処理

これまで見てきたように同一言語の2つのパラレルコーパス，K_J と K_E は同一内容を示しているにもかかわらず，用いられている表現形式はかなり異なる．両コーパスともプロの翻訳家によって作成されたコーパスであるため，極端に不自然な文は存在しないといえる．そのため，この2つのパラレルコーパスは，同義表現を得るための言語資源として非常に適しているといえる．

翻訳を介しての同義表現の獲得に関しては，これまで検討されており（たとえば，(Barzilay and McKeown 2001) など），その可能性が示されている．この場合の，言語資源としては，ある言語 A における文書 X とそれの言語Bへの翻訳文書 Y だけでは不十分である．この場合は，Y を翻訳した人間とは別の者による X から言語 B への翻訳 Z も存在してはじめて，Y と Z を比較することにより同義表現の獲得が可能となる．これに対して，本研究では，言語 A と B のパ

ラレルコーパス$C(A)$と$C(B)$をそれぞれ別の言語Kのコーパス$C(K_a)$と$C(K_b)$に翻訳しており，同一言語間の複数の翻訳を用いる状況とは異なる．しかも，翻訳の原言語が，本研究では，英語と日本語であり，その構造は大きく異なる．したがって，本研究で用いたコーパスそのものを同義表現獲得に有効に用いることができる．

たとえば，図2に示すように日本語原文に対する韓国語訳は，構文ならびに語彙の面で原文と大きな違いはない．一方,「食欲がない」に対する英訳は “I have no appetite.” であり，それを韓国語に訳すと「밥 생각이 없어요. （ご飯の思いがない.)」となる．これは韓国語では日常よく使われる表現である．しかし，ここで韓日機械翻訳を考えると，韓国語では自然な表現であっても，その翻訳が “ご飯の思いがない．” となることは好ましくない．これは日韓の翻訳においても同様である．そこで，これらのコーパスを用いてあらかじめ換言知識を抽出しておき，換言器を構成する．この換言器を用いることによって「ご飯の思いがない.」に対応する原文を「食欲がない.」に対応する原文に換言することができる．その結果より自然な訳文を得ることができる．

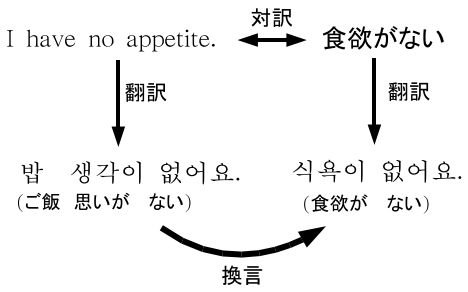

図 2　同義表現を用いた換言

近年，大量の二言語パラレルコーパスを用いる統計翻訳が注目されている．本研究で用いた日本語コーパス(Jとする)と英語コーパス(Eとする)ならびにそれらの韓国語翻訳であるK_JとK_Eを用いると，日-英-韓の間で翻訳機を構成することができる．今，韓国語のコーパスとしてK_Jのみを用いて韓国語を目的言語とする翻訳機を考える．この場合，日-韓の翻訳機では，自然な翻訳を期待できるが，英-韓の翻訳機では，翻訳モデル内の対応付けが，より複雑になることが予想され，翻訳精度が低下する可能性がある．そのため，常にK_Jを使用することが推奨されるわけではない．むしろ，統計翻訳のようなコーパスに基づいた機械翻訳機の場合，使用するコーパスが直訳調の対応になっている方が，計算機にとっては，対応がとりやすく処理しやすいと言える．したがって，ひとつのモデルとして，翻訳に使用するコーパスは直訳調のものを用いて(たとえば，英-韓の翻訳機の場合，E-K_Eを使用する)翻訳機を構成し，翻訳前/後の文を同義表現知識を用いて換言するものが考えられる．

5 結論

日英パラレルコーパスの日本語と英語それぞれを原言語として翻訳した2種類の韓国語旅行会話コーパスを用いて，原言語が翻訳に及ぼす影響についていくつかの言語現象を分析した．要約すると，文法および語彙において非常に類似している日本語と，それらが相当異なる英語それぞれからの翻訳では，原言語の違いが翻訳に多大な影響を与えている事実を示すことができた．これは人間の翻訳者においても機械翻訳においても同じことだと考える．今後はこのような言語差を利用した同義表現の抽出について詳しく検討する予定である．

謝辞

本研究は総務省の研究委託「携帯電話等を用いた多言語自動翻訳システム」により実施したものである．

参考文献

Barzilay, R. and McKeown, K. R. (2001). “Extracting Paraphrases from a Parallel Corpus.” In *Proceedings of the 39th Annual Meeting of the Association for Com putational Linguistics*, pp. 50–57.

張元哉 (2000). “19世紀末の韓国語における日本製漢語–日韓同形漢語の視点から–.” 日本語科学, **8**, 76–95.

Culy, C. and Riehemann, S. Z. (2003). “The limits of N-gram translation evaluation metrics.” In *MT Summit IX* New Orleans.

Downing, P. (1996). *Numeral Classifier Systems, the case of Japanese*. John Benjamins, Amsterdam.

Kaiser, S., Ichikawa, Y., Kobayashi, N., and Yamamoto, H. (2001). *Japanese: A Comprehensive Grammar*. Routledge.

Lee, Y.-O. (1999). “The Difference in Subject Choice between Korean and English.” In *English Education in the Era of Information*. Chungnam National University, Kwangju, Korea.

Lehmann, M. (2000). “String::Similarity.” Perl Module (`http://cpan.org/`). (v0.02).

Myers, E. (1986). “An O(ND) difference algorithm and its variations.” *Algorithmica*, **1** (2), 251–266.

Ohtake, K. and Yamamoto, K. (2001). “Paraphrasing Honorifics.” In *Workshop Proceedings of Automatic Paraphrasing: Theories and Applications (NLPRS2001 Post-Conference Workshop)*, pp. 13–20 Tokyo.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). "Bleu: a Method for Automatic Evaluation of Machine Translation." Tech. rep., IBM Research Division Tohmas J. Watason Research Center. IBM Research Report RC22176(W0109-022).

Shibatani, M. (1990). *The languages of Japan*. Cambridge Language Surveys. Cambridge University Press.

Sohn, H. (1999). *The Korean Language*. Cambridge Language Surveys. Cambridge University Press.

Takezawa, T., Shirai, S., and Ooyama, Y. (2001). "Characteristics of Colloquial Expressions in a Bilingual Travel Conversation Corpus." In *19th International Conference on Computer Processing of Oriental Languages: ICCPOL-2001*, pp. 384–389 Seoul.

Unterbeck, B. (1994). "Korean Classifiers." In Kim-Renaud, Y.-K. (Ed.), *Theoretical Issues in Korean Linguistics*, pp. 367–385. CSLI.

渡辺吉鎔, 鈴木孝夫 (1981). 朝鮮語のすすめ. 講談社.

付録

A コーパス中の例文

本研究にて使用したBTECの一部を表8に示す.

表 8 本研究にて用いたBTECの一部

韓国語	英語/日本語	類似度
네?	Yes?	0.1429
용건은?	ご用向きは.	
저에게 질문해 주세요.	Ask me.	0.2745
나한테 물어 봐.	私に聞いて.	
내 표가 보이지 않는군요.	I lost my ticket.	0.3438
표를 잃어 버렸습니다.	切符をなくしてしまいました.	
담배 좀 피려고 하는데 괜찮겠어요?	Do you mind if I smoke?	0.4156
담배를 피워도 됩니까?	たばこを吸ってもかまいませんか.	
그 여자는 안경을 끼지 않았어요.	She wasn't wearing glasses.	0.5238
안경을 쓰고 있지 않았습니다.	眼鏡をかけていませんでした.	
그가 머리를 다쳤어요.	He has hurt his head.	0.6984
그는 머리를 다쳤습니다.	彼は頭にケガをしました.	
이제 됐습니까?	Is that Okay?	0.7907
이러면 됐습니까?	これでいいですか.	
선물 가게 위치가 어디입니까?	Where is the gift shop?	0.8056
선물가게는 어디입니까?	ギフトショップはどこですか.	
블랙으로 주십시오.	Black, please.	0.9286
블랙으로 해 주십시오.	ブラックにしてください.	

略歴

白 京姫（ペク キョンヒ）： 1993年慶應義塾大学大学院社会学研究科修士課程教育学専攻修了．1998年同大大学院社会学研究科後期博士課程教育学専攻単位取得退学．1999年CSLI, Stanford大学客員研究員．2000年〜2005年(株)国際電気通信基礎技術研究所（ATR）音声言語コミュニケーション研究所研究員．韓国語の自然言語処理，および助数詞の解析・生成等の研究に従事．2005年より翻訳業．ACL, ALS各会員．email: paikbond@gmail.com

大竹 清敬（おおたけ きよのり）： 2001年豊橋技術科学大学大学院工学研究科博士後期課程電子・情報工学専攻修了．博士(工学)．2001年より(株)国際電気通信基礎技術研究所(ATR)音声言語コミュニケーション研究所研究員．言語表現変換技術とその応用（要約，翻訳など），中国語処理，コーパス利用のための技術（表記のゆれ処理など），形態素解析と単語解析のための辞書構築などに興味がある．言語処理学会，人工知能学会，情報処理学会各会員.

e-mail: otake@fw.ipsj.or.jp

Francis Bond（フランシス ボンド）： 1988年 B.A. (University of Queensland)．1990年 B.E. (Hons) (同大学)．1991年 日本電信電話株式会社入社．以来，計算機言語学，自然言語処理，特に機械翻訳の研究に従事．1999年 CSLI, Stanford大学客員研究員．2001年 Ph.D. (University of Queensland)．2005年 3ヶ月間Oslo大学招聘研究員．現在，NTTコミュニケーション科学基礎研究所主任研究員．著書「Translating the Untranslatable」, CSLI Publications にて日英機械翻訳における数・冠詞の問題を扱う．ACL, ALS, 言語処理学会各会員．email: bond@cslab.kecl.ntt.co.jp

山本 和英（やまもと かずひで）： 1996年豊橋技術科学大学大学院工学研究科博士後期課程システム情報工学専攻修了．博士(工学)．1996年〜2005年(株)国際電気通信基礎技術研究所(ATR)研究員(2002年〜2005年客員研究員)．1998年中国科学院自動化研究所国外訪問学者．2002年より長岡技術科学大学電気系，現在助教授．言語表現加工技術(要約，換言，翻訳)，アジア言語処理(中国語，韓国語など)，言語処理技術を活用したテキストマイニングなどに興味がある．言語処理学会，人工知能学会，情報処理学会，ACL 各会員．
e-mail: yamamoto@fw.ipsj.or.jp

(2004 年 9 月 30 日 受付)
(2005 年 1 月 14 日 再受付)
(2005 年 4 月 23 日 採録)